STATISTICS AND THE MILITARY OFFICER IN COMBAT DEVELOPMENTS EXPERIMENTATION

by

David Warren Mikkelson

UNITED STATES
NAVAL POSTGRADUATE SCHOOL

THESIS

STATISTICS AND THE MILITARY OFFICER IN COMBAT DEVELOPMENTS EXPERIMENTATION

by

David Warren Mikkelson

December 1968

STATISTICS AND THE MILITARY OFFICER IN COMBAT DEVELOPMENTS EXPERIMENTATION

by

David Warren Mikkelson
Captain, United States Army
B. S., South Dakota School of Mines and Technology, 1962

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
December 1968

ABSTRACT

The duty performance of military officers whose duties are the planning, conduct, analysis, and evaluation of field experimentation can be improved through a better understanding of experimental statistics. The role of statistics in the field experimentation conducted by the U. S. Army Combat Developments Command Experimentation Center typifies the role of statistics in military field experimentation. Selected officers of USACDCEC were surveyed to determine their understanding of some of the more important concepts of experimental statistics. The survey results indicate that most of these officers lack a basic knowledge of experimental statistics. Based on insights gained from the survey, statistical training of certain USACDCEC officers is recommended. Statistical concepts not well understood by the surveyed officers are defined and discussed. A field experiment conducted by USACDCEC is used to exemplify the applications of statistical techniques and the use of measures of performance in field experimentation.

TABLE OF CONTENTS

CHAPTER		PAGE
I.	INTRODUCTION	8
II.	A SURVEY OF USACDCEC OFFICERS UNDERSTANDING OF STATISTICS	13
	Discussion of Survey Results	15
	General Information	15
	Essay Questions	17
	Matching Section	18
	Conclusions Drawn from Survey	22
	Recommendations	22
III.	SOME BASIC CONCEPTS OF STATISTICS	25
	The Role of Statistics in Experimentation	25
	Hypothesis Testing	27
	Type I and Type II Error	28
	Statistical Significance	32
	Test Statistic	34
	Replication	35
	The Relationships Among Type I Error, Type II Error, and Replication	38
	Confidence Intervals	41
	Suggested Readings	45
IV.	DISCUSSION OF AN EXPERIMENT	48
	The Experiment	49
	Measures of Fire Effectiveness	49
	Resources	50
	Firing Courses and Traverse Speeds	51
	Target Arrays	51
	Fire Discipline	52

PAGE

Subject Personnel - - - - - - - - - - - - - - - - - - - 52
Replication of Trials - - - - - - - - - - - - - - - - - 54

Findings of the Experiment - - - - - - - - - - - - - - 55
Accuracy - 57
Time to Obtain a First Hit - - - - - - - - - - - - - - 62
Index of Target Area Hits - - - - - - - - - - - - - - 67
Volume of Effective Fire - - - - - - - - - - - - - - - 69

Conclusion of the Experiment's Report - - - - - - - - 72

Summary of Comments About the Experiment - - - - - 75

BIBLIOGRAPHY - 78

APPENDIX A. Statistics Survey Questionnaire - - - - - - - 79

APPENDIX B. Summary of and Comments on Survey Results - 87

LIST OF TABLES

TABLE		PAGE
I.	Questionnaire Response by USACDCEC Element - -	15
II.	Subject Performance on Matching Section - - - - -	20
III.	Type of Errors Committed - - - - - - - - - - - - -	30
IV.	Relation Between Replications and Type I and II Errors -	40
V.	Relationship Among Degree of Confidence, Size of Confidence Interval, and Number of Replications - -	44
VI.	Summary of Trial Data - - - - - - - - - - - - - - -	56
VII.	Average Traverse Time and Hits Per 50-Meter Segment Per Individual Firer - - - - - - - - - - - -	63
VIII.	Average Rate of Fire of Individual Firers - - - - -	66
IX.	Average Hits on Sensor Targets Per Second of Individual Firers - - - - - - - - - - - - - - - - - -	70
X.	Distribution of Ratings Received by Question - - -	89
XI.	Types of Replies to Selected Terms - - - - - - - -	93

LIST OF FIGURES

FIGURE		PAGE
1.	Number of Subjects Correctly Identifying a Specific Term -	21

ACKNOWLEDGMENTS

The assistance and cooperation of the United States Combat Developments Command Experimentation Command and its officers, especially the efforts of Lieutenant Colonel Ernest Phillips, are greatly appreciated. Sincere thanks are also expressed to the personnel of Litton Scientific Support Laboratory, Fort Ord, California, and the Army and Marine students in the Operations Analysis program at the U. S. Naval Postgraduate School who contributed to the statistical survey which is the cornerstone of this thesis.

CHAPTER I

INTRODUCTION

This thesis is directed primarily toward the military officer who is involved with field experimentation. It is intended to provide a brief orientation on the meanings and relationships of some of the more important concepts of experimental statistics, to provide examples of applications of statistics in field experimentation, and to demonstrate the need for the military officer to have a basic knowledge of statistics.

The term statistics is recognized to have a dual meaning dependent upon the context in which the term is used. One definition is that statistics are numerical results obtained by arithmetic operations on numerical data, while the other definition is that statistics is the science of methods and procedures used to obtain numerical results (statistics), to estimate the reliability of the results, and to draw inferences from results. The latter definition of statistics is the context in which the term is used in most instances throughout this thesis.

The role of statistics in military field experimentation is exemplified by its role in the applications of field experimentation by the United States Army. In order to assist its efforts in the formulation of new doctrine, organizations, and materiel objectives and requirements for the Army Combat Development Program, the

United States Army has established the U. S. Army Combat Developments Command Experimentation Center at Fort Ord, California. The mission of USACDCEC is to "conduct scientific field experimentation that:

(1) Develops and provides experimentation derived data as input for the models, simulations, or war games used by USACDC* agencies and institutes in their scientific analysis and evaluation of various alternative solutions to combat development actions.

(2) As directed, tests, analyzes and provides experimentally derived data on developmental options created by USACDC agencies and institutes.

(3) Examines for validity basic rationale used in the scientific analysis actions of USACDC agencies and institutes.

(4) Verifies, through field experimentation, recommended solutions for operational concepts, materiel requirements and organizational structures."[1]

The USACDCEC mission is accomplished by the joint effort of the Army and a contracted civilian scientific support laboratory.

The basic postulate of this thesis is that Army officers, whose duty assignments at USACDCEC require them to participate in the planning, conduct, analysis and evaluation of field experimentation, need to have a good understanding of the basic concepts of experimental statistics. An understanding of statistics by these officers need not be a detailed theoretical knowledge of the mathematical and probabalistic aspects of statistics, but their knowledge of

---

*USACDC - United States Army Combat Developments Command

[1]"Experimentation Manual", UNITED STATES ARMY COMBAT DEVELOPMENTS COMMAND EXPERIMENTATION COMMAND (Fort Ord, California, 1968), p. 6. (Mimeographed.)

statistics should encompass an understanding of the general concepts of experimental statistics and its application to the empirical aspects of military operations research and quantitative decision making. Even though the role of the civilian scientist in military field experimentation is to provide scientific expertise, the justification and control of the resources used in field experimentation are the responsibility of the Army, as is the final responsibility for the content of reports containing findings and conclusions based on field experimentation. At least in a general manner, the Army officer should understand the statistical concepts underlying the civilian scientist's recommendations for the statistical design and analysis of an experiment. To best understand the scientist's recommendations, the Army officer should have a rudimentary knowledge of the statistical methodology employed by the scientist.

The first step of this thesis effort was to establish whether or not the "typical" officer at USACDCEC was or was not lacking in his understanding of statistics. Permission was obtained to survey the comprehension of statistics of officers assigned to USACDCEC whose duty assignments require them to participate in the statistical planning and analysis of field experiments. Chapter II discusses the formulation and administration of the survey and elaborates on the results of the survey.

The results of the survey clearly indicate that most of the surveyed officers did not comprehend many of the concepts that are

basic to a good understanding of statistics. To overcome the deficiency in the statistical education of many of the officers, it is proposed that an educational program in statistics be developed at USACDCEC. Oriented specifically to combat developments field experimentation, such a program could provide the USACDCEC officer with sufficient statistical training to substantially increase his ability to understand the applications of statistics and, thereby, make him a more effective member of the soldier-scientist team.

The second step of this thesis effort was to provide to the non-statistician an explanation of some of the concepts that are a foundation of experimental statistics. The topics discussed in Chapter III are by no means inclusive of all the ideas that should be learned to achieve a basic understanding of statistics, but they are considered to be concepts which are especially essential for the military officer who is involved with field experimentation to understand. Chapter IV discusses the application of statistical techniques and the uses of measures of performance in an actual experiment conducted by USACDCEC in 1964.

CHAPTER II

A SURVEY OF USACDCEC OFFICERS' UNDERSTANDING OF EXPERIMENTAL STATISTICS

Does a problem really exist? This is one of the key questions that must be answered when examining any situation which is suspected of containing problem areas.

Do USACDCEC officers, who should have a good grasp of the concepts of experimental statistics, really have a basic understanding of the language of statistics? To provide information from which to base an answer to the above question, a survey questionnaire was developed to sample the attitudes toward and understanding of statistics from officers of selected elements of USACDCEC.

The elements of USACDCEC selected for the survey were the Field Experimentation Division of the G-3 Staff and Project Teams I, II, III, IV, and V. The responsibilities for planning, conduct, and analysis of USACDCEC field experiments is primarily with these elements of USACDCEC. With the exception of a few administrative positions, the officer positions in these six elements are positions that should contain officers who understand the basic elements of statistics.

The survey questionnaire was not hastily conceived. It underwent revision before reaching its final form as it appears in Appendix A. Two Army officers, formerly assigned to USACDCEC and now

first year students in the Operations Research/Systems Analysis graduate program at the Naval Postgraduate School, completed the questionnaire before its final revision. Their comments on the clarity and understandability of the questionnaire were helpful in producing the final questionnaire format and wording.

The questionnaire was distributed to the surveyed USACDCEC elements on 12 July 1968 and returned by 23 July 1968. Over a week was allowed to provide sufficient time for subjects to complete the questionnaire.

The total assigned strength of the six surveyed elements was seventy-one officers. Officers in the six elements whose duties assignments were not directly connected with the planning, conduct, analysis and review of experiments were exempted from the survey.

Thirty-five questionnaires were completed and returned. Some of the officers assigned to the surveyed elements were absent on leave or temporary duty. Considering that administrative personnel were exempted from the survey, the response by thirty-five officers represents over 50 per cent of the population sampled. The response by element is indicated in Table I.

TABLE I

QUESTIONNAIRE RESPONSE BY USACDCEC ELEMENT

Element	Questionnaires Completed and Returned
G-3 FED	4
Team I	6
Team II	6
Team III	9
Team IV	5
Team V	5
Total	35

The questionnaire is composed of three sections which required a subject's response. The first section requests general information about each subject. The second section is comprised of essay questions, the response to which should provide a feeling of the subject's attitude about statistics, the field experimentation mission, and the interface with the civilian scientist. The third section is a list of terms and phrases and definitions which are representative of the language of statistics that a statistically sophisticated officer should know and understand. The subjects were requested to match each term with the best definition of the term.

I. DISCUSSION OF SURVEY RESULTS

General Information

Appendix B presents a summary of the results of the general information section of the survey.

The survey sample is considered representative of the types of officers and their respective qualifications that will be available to staff the surveyed USACDCEC elements in the next few years. At present, the availability of officers with strong backgrounds in mathematics and operations research is insufficient to satisfy the growing number of Army job positions requiring these scientific skills. No relief from this critical shortage of specially trained officers is in sight for the next few years even though the Army has significantly accelerated the training of officers in these skills.

Twenty-one of the thirty-five officers replied that a better knowledge of statistics would definitely assist them to increase the effectiveness of their duty performance. Six replied that it would be of marginal value; four indicated that they didn't think it would help; and three replied that it definitely would not help. With the exception of one Major, all of the officers who felt a knowledge of statistics would not help in their job performance were Lieutenants. One Lieutenant Colonel did not reply directly but reasonably stated, "without having been exposed to the subject [statistics], I cannot judge what its value might be".

Fifteen officers indicated that they have had at least some type of formal instruction in statistics. Twelve of these fifteen were part of the twenty-seven officers who felt a better statistical knowledge would be at least of some value to them. This is a clear indication from those who have had some exposure to statistics that knowing more

about statistics would assist them in their job performance.

From the length of time that the subjects took to complete the questionnaire, one can infer that most of the subjects put forth a conscientious effort in the survey. Only three subjects indicated that they took less than one hour; four subjects did not indicate their questionnaire completion time.

Essay Questions

The nature of the responses to the fourth and sixth essay questions are especially noteworthy.

Question 4. "Assume we are comparing the performance of two configurationally different platoons in a certain measure of effectiveness. What is the meaning of the statement 'there is a significant difference at the 5 per cent level between platoons' or 'the significance level for a difference between platoons is 5 per cent'?" Twelve officers felt that the statements meant that there is a 5 per cent difference in platoon effectiveness, e.g., as one officer stated, "Differences in performance of 5 per cent or greater are significant. Differences in performance between the two organizations of less than 5 per cent are not significant." The level of statistical significance is not synonomous to operational significance. Only one officer correctly and specifically identified the use of Type I experimental error and its implications to the statements, i.e., the probability of committing a Type I error, concluding a difference exists when, in fact, there is no difference, is 5 per cent.

Question 6. "Would a reference on experimental statistics, written specifically for officers with little or no formal statistical training, be of some value to you? If yes, in what way?" Twenty-three officers responded with "yes" and one replied with "possibly". Eight officers replied with "no". Two of the eight felt that the existing literature is sufficient and one did not want a reference but rather wanted classroom instruction. Three officers gave no reply. The twenty-three "yes" responses included a general agreement that a layman's statistics reference would help to increase the military officer's ability to work more effectively in his job and to understand the recommendations of civilian scientists.

A discussion of the results of the other essay questions is contained in Appendix B.

Matching Section

As a check on the validity of the matching section, ten Army and Marine officers, second year students in the Operations Research/Systems Analysis program at the Naval Postgraduate School, volunteered to complete the matching section of the questionnaire. The ten student officers had finished six quarters of the OR/SA program which included two courses in probability theory, a course in statistics, and a course in methods of combat developments experimentation. They completed the matching section on 8 July 1968. By virtue of their recent statistical training, the student officers should have a good understanding of statistical terms. Thus, a good performance

by the student officers would validate the matching section's answerability as well as provide a basis for comparison of the USACDCEC officers' performance on the matching section.

Table II reflects the overall performance of officers completing the matching section.

In a comparison of performance, the inference that is drawn is that USACDCEC officers failed to score better because they did not understand many of the terms and definitions that they were asked to match and not because of ambiguities in the terms and definitions presented to them. This is not too surprising in view of the fact that only fifteen USACDCEC officers indicated they had at least some training in statistics. To further emphasize this inference, note that USACDCEC officers scored 231 correct out of 700 (35 x 20), or 33.0 per cent, and averaged 6.6 correct out of 20 per officer. USNPGS student officers scored 185 correct out of 200 (10 x 20), or 92.5 per cent, and averaged 18.5 correct out of 20 per officer. The worst performance of a student subject was as good as the best performance of a USACDCEC subject, 15 out of 20 correct.

The frequency of correct answers by USACDCEC officers on each term of the matching section is represented in Figure 1.

Only six of twenty terms in the matching section received over 50 per cent correct responses from USACDCEC officers. A term-by-term discussion of responses is found in Appendix B beginning on page 87.

TABLE II

SUBJECT PERFORMANCE ON MATCHING SECTION

Number of Correct Answers	Number of Subjects USACDCEC Officers	USNPGS Students
20 (Perfect)		3
19		3
18		2
17		1
16		
15	1	1
14		
13	2	
12		
11	1	
10	2	
9	2	
8	4	
7	6	
6	5	
5	3	
4	2	
3	4	
2	1	
1		
0	2	
TOTAL	35	10

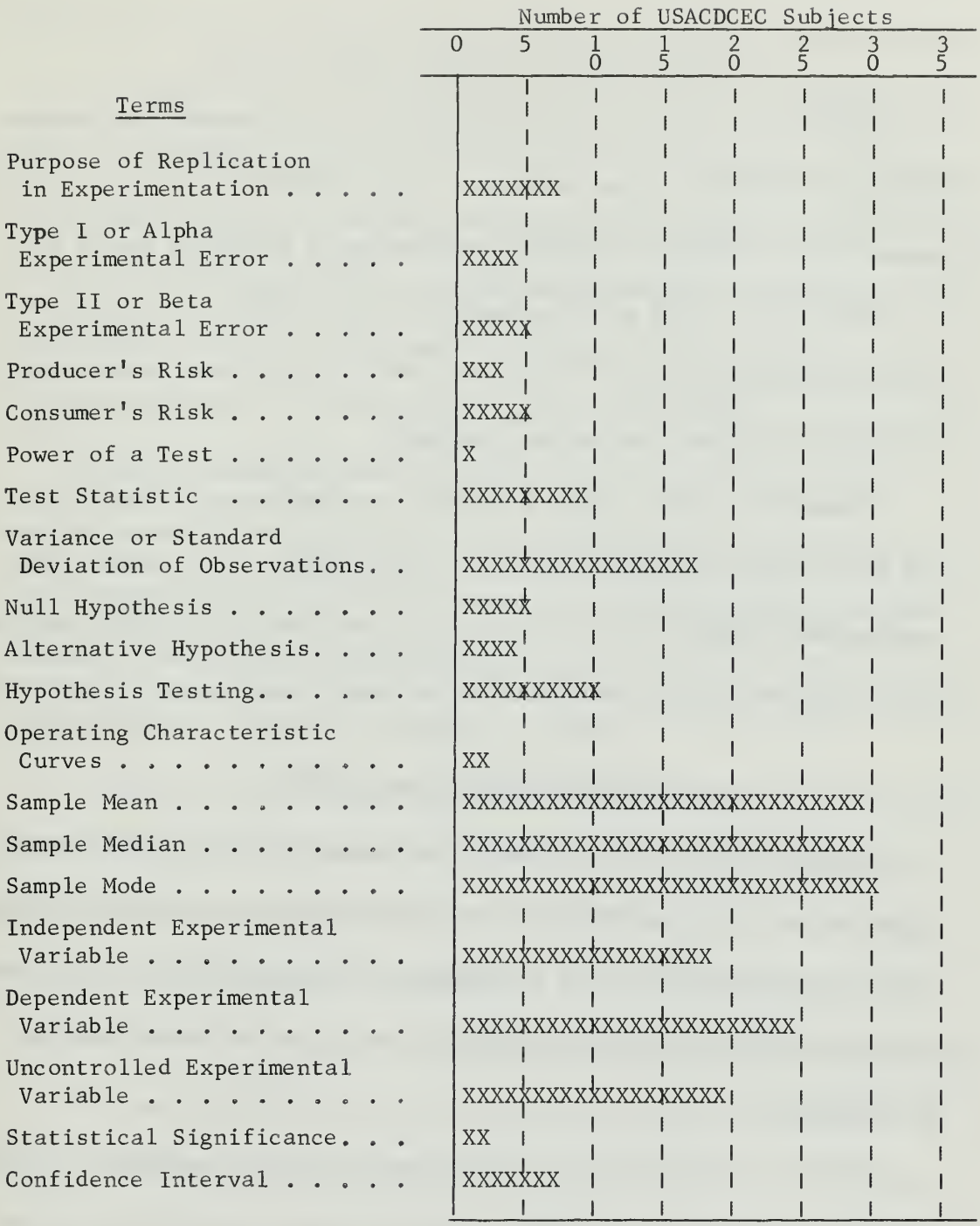

FIGURE 1

NUMBER OF SUBJECTS CORRECTLY IDENTIFYING A SPECIFIC TERM

II. CONCLUSIONS DRAWN FROM SURVEY

The survey sample is considered to be a representative sample of those USACDCEC officers whose duty performance involves routine contact with situations in which a knowledge of statistics would be very beneficial. With a few exceptions, these officers do not have a good understanding of statistics; but, they are desirous of obtaining an education in basic statistical concepts.

It should be made clear, however, that the intent of the survey was not to embarrass the officers of USACDCEC. The results and conclusions should in no way be used to infer that USACDCEC officers do not understand their functions in the field experimentation mission.

III. RECOMMENDATIONS

Recognizing that the availability of officers with special training in mathematics and operations research is limited, USACDCEC could find it very beneficial to the accomplishment of its mission to provide some type of training in experimental statistics for the officers of the command.

Formal classroom-type instruction within the command is recommended to achieve best results. Classroom presentations could be designed and tailored to the peculiarities of the Army combat developments field experimentation mission. The content of instruction should be oriented towards a management-level treatment of the concepts and methodology of the application of statistics to field

experimentation. The presentation of detailed mathematical derivations of specific statistical techniques should not be included. However, a "quickie" type course of only a few hours should be avoided if a longer course is feasible. With careful attention to course content, a course in the range of 35 to 40 hours beginning with a block of 8 to 10 hours of a limited treatment of probability theory should be of sufficient length to provide a good basic understanding of statistics.

If the establishment of "in house" formal instruction seems overly ambitious or too costly in the consumption of duty time, perhaps selected members of the command could be permitted to audit classes at the Naval Postgraduate School in which probability theory and statistics are part of the course content.

As a final resort, a self-teaching correspondence type course in statistics could be developed for those officers who recognize and want to overcome their limited ability to use statistics and to communicate in the language of statistics. Each officer interested in self-education in statistics could proceed at his own speed to the level of understanding that he deems sufficient for him to accomplish his duties most effectively. Professional statisticians of the Scientific Support Laboratory could be called upon for individual assistance as required as each officer progresses with his personal self-education program.

USACDCEC should consider the possibility of interesting Army students in the Operations Analysis program of the U. S. Naval

Postgraduate School in the potential thesis areas that the establishment of a program of instruction in statistics would provide. For example, a thesis effort could develop a detailed program of instruction on an 8 to 10 hour block of instruction to introduce probability theory as part of a course in experimental statistics.

CHAPTER III

SOME BASIC CONCEPTS OF STATISTICS

The purpose of this Chapter is to explain some of the statistical concepts which are contained in the essay and matching sections of the survey questionnaire. The discussion of statistical terms and phrases of the matching section of the questionnaire will be limited to those terms and phrases that were not correctly identified with a frequency of over 50 per cent by the USACDCEC survey subjects. Thus, the terms sample mean, median, and mode and experiment variables will not be discussed.

The discussion will be kept brief and simple for the reader with little knowledge of mathematics and probability theory. It is assumed, however, that the reader does have at least a notion of what a probabalistic statement means, e. g., that he understands that an event having a probability of 95 per cent of occurring has a "good" chance of a realization and that an event having a probability of 5 per cent of occurring has relatively "poor" chance of realization.

The Role of Statistics in Experimentation

Statistics provides the mathematical basis of the design and analysis of an experiment. After the recognition and consideration of the constraints of resource availability, time, money, men, and materiel, a statistical plan for the experiment is developed. The plan designates the number of replications on the experiment to be

conducted, how data is to be reduced and analyzed, the statistical tests to be employed, and the form of the inferential statements that can be made about the results of the experiment.

The importance of experimental design is well stated in the United States Army Combat Developments Command's "Methodology Notebook for Action Officers"[2] which says:

> The commitment of resources essential to the conduct of a field experiment must be preceded by a meticulously conceived and developed experimental plan or design. Failure to predetermine the total methodology--to include precise details on the handling of many deviations that might occur during the actual experiment--will likely result in a failure to attain valid experimental results. The scientific planning effort expended before the initiation of the actual experiment is perhaps the most important aspect of successful field experimentation....

The two general areas of experimentation which employ statistical techniques are hypothesis testing and estimation. Hypothesis testing may be employed to make comparisons, e.g., determining whether an observed difference between two units in a measure of performance is significant in a statistical sense. Estimation is the determination of an unknown characteristic of the experimental unit or subject, e.g., determining the operational hit probability of a weapon system.

Statistics provides a mathematical structure to scientific method and logic in attempting to answer combat development questions by

---

[2] May 1967, Chap 6, para 6a.

the conduct of field experimentation. Providing an objective statistical basis upon which decisions can be made is the purpose of experimentation. This objectivity in experimentation is the primary difference from subjectively derived results from less rigorously controlled troop tests in which information is non-mathematically analyzed.

Hypothesis Testing

Webster's New Collegiate Dictionary defines a hypothesis as "a tentative theory or supposition provisionally adapted to explain certain facts and to guide in the investigation of others". Note the use of the adjective "tentative" and the adverb "provisionally" in the definition. In the process of applying the scientific method to military field experimentation, a hypothesis is formulated on which an attempt will be made to reject that hypothesis by the results of the experiment. It is one of the basic tenets of the scientific method that hypotheses are never proved but they can be shown to be quite unlikely in the light of empirical evidence. In statistical terminology this hypothesis is called the null hypothesis since the comparison made by the null hypothesis is frequently one of equality, i.e., there is no difference or a null difference in the comparison.

In the example of comparing two platoons in a measure of performance (essay question 4 of the survey), the null hypothesis would be stated as "the average performance of the two platoons in the measure of performance are the same". An alternative to the null

hypothesis, appropriately called the alternative hypothesis, could be stated as "the average performance of the two platoons in the measure of performance are not the same". Rejection of the null hypothesis by the results of an experiment infers that there is a difference in the average performance of the two platoons. However, a failure to reject the null hypothesis does not always infer that the average performance of the platoons are the same. In certain cases, the results of an experiment can only legitimately be called inconclusive, since an actual difference may exist and the hypothesis test used in the experiment may not be strong enough to detect that difference.

Type I and Type II Error

Because an experiment normally tests samples from a large population in order to answer questions about certain characteristics of the population, an element of chance is always involved in hypothesis testing. The true nature of a population's characteristic can be exactly determined only if all items of the population are tested with complete accuracy. In such a case, no chance is involved since the characteristic is exactly determined.

As an example, suppose it is desired to know if there is a difference in the mean (average) gas mileage between 1968 Ford and Plymouth automobiles with comparable engines. One method that could be used, though highly infeasible, would be to accurately test every 1968 Ford and Plymouth; the average gas mileage for each

make of automobile is then completely determined with certainty. A feasible method would be to select and test a sample number of automobiles from the entire population of 1968 Fords and Plymouths and to compare the sample mean gas mileage of each make of automobile; but, there is a degree of uncertainty about whether or not the sample means really reflect the true, but unknown, average gas mileage of each make. In actuality, either there is or there is not a difference in the average gas mileage. Only by testing the entire population of each automobile can one know for sure if a difference exists.

The null hypothesis for this situation can be stated as "there is no difference in mean gas mileage". The alternative can be stated as "there is a difference in mean gas mileage". If the sampling method is used, there are chances of reaching the wrong conclusion. These chances are referred to as the probabilities of a Type I and Type II error.

Type I error is the rejection of the null hypothesis when, in fact, it is true; while Type II error is the acceptance of the null hypothesis when, in fact, it is not true. For example, a Type I error is committed if the experiment leads to the conclusion that there is a difference in average gas mileage when the truth, unknown to the experimenter (or anyone), is that there is no difference. Conversely, a Type II error is committed if the experiment leads to the conclusion that there is no difference in average gas mileage

when the truth is that there is a difference. Table III summarizes the situations in which a Type I or Type II error is committed.

TABLE III

TYPE OF ERRORS COMMITTED

If the null hypothesis is:	When, in fact (but unknown) the null hypothesis is:	
	True	False
Accepted	No Error	Type II Error
Rejected	Type I Error	No Error

The terms producer's and consumer's risk have arisen in product quality assurance applications of industrial statistics. The probability of a Type I error has been referred to as producer's risk while the probability of a Type II error has been referred to as consumer's risk.

To motivate the inferences of producer's and consumer's risk, consider the following example. An ammunition manufacturer is producing a particular caliber of small arms ammunition for the Army. A specification for the ammunition requires that a certain muzzle velocity be attained from the ammunition to insure proper functioning of the using weapon. Before accepting a shipment of ammunition from the manufacturer, the Army desires to test samples of the ammunition to determine if the muzzle velocity specification has been met. The null hypothesis is "the ammunition produces the desired muzzle velocity".

The probability of rejecting the null hypothesis, when it is true, is the chance or risk the manufacturer must take, hence it is called the producer's risk. The probability of accepting the null hypothesis, when it is false, is the chance or risk the Army must take, hence it is called the consumer's risk. The manufacturer risks a refusal of a shipment of truly acceptable ammunition and the accompanying monetary loss. The Army risks the receival of a shipment of truly unacceptable ammunition which could lead to dire circumstances on the battlefield.

When conducting field experimentation, the Army is in a sense both the consumer and the producer. Rejection of truly good combat development concepts or acceptance of truly poor concepts that are being experimentally tested unquestionably can have a detrimental impact on the Army's ability to perform its missions in an optimal manner. The point is - attention must be given to the implications of both Type I and Type II error in combat development experimentation.

There is an inverse, and sometimes troublesome, relationship between Type I and Type II error probabilities. For a given experiment with a fixed number of replications, if Type I error probability is permitted to decrease, Type II error probability will correspondingly increase; conversely, if Type I error probability is permitted to increase, Type II error probability will correspondingly decrease.

As an example, an experiment could yield results* whereby, if the probability of a Type I error is 1 per cent, the corresponding Type II error probability is 50 per cent. Suppose the decision maker desires to use a larger Type I error probability of 5 per cent instead of 1 per cent: Then, the corresponding Type II error probability will decrease to 20 per cent.[3] In this example, a large decrease in the probability of a Type II error results from a relatively small increase in the probability of a Type I error.

The determination of the appropriate trade-off between the magnitudes of Type I and Type II error probabilities is dependent on the costs associated with committing Type I and Type II errors. This trade-off determination is a problem which is attacked by the methodologies of statistical decision theory.**

Statistical Significance

The results of a hypothesis test are statistically significant if the test leads to the rejection of the null hypothesis. The significance level of a hypothesis test is that level of Type I error probability at

---

* Hypothetical results - $(m-m_o)$/standard deviation = 1, n = 10.

** An adequate discussion of statistical decision theory is beyond the scope of this thesis. The reader is referred to the Suggested Readings section at the end of this chapter.

3 S. Eherenfeld and S. B. Littauer, Introduction to Statistical Methods (New York: McGraw-Hill, 1964), Figures 7.7 and 7.8, p. 272.

which it is statistically permissible to reject the null hypothesis. To say a test is significant at the 5 per cent level means that the probability of the rejection of a null hypothesis which is in reality true is 5 per cent. Or in another sense, one could say that the probability of accepting a null hypothesis which is really true is 95 per cent.

Significance levels of low probability are normally used in analyzing experimental data to give the decision maker a high degree of confidence that he is not committing a Type I error. Thus, in the ammunition example, the manufacturer would desire a very low significance level for the analysis of test data on his ammunition. And, in fairness to the manufacturer, he should be granted a reasonably low significance level acceptable to both him and the Army, assuming the Army can maintain its desired level of Type II error probability.

Statistical significance should indicate operational significance. The statistical design of an experiment should be such that if the tested items are found to be statistically different in a measure of performance, then the tested items should also be operationally different in that same measure of performance. When planning an experiment, the decision maker should subjectively determine what is operationally significant in a measure of performance. Then the statistician can statistically design the experiment to reveal that operational significance as statistically significant in the analysis

of experimental data, if such operational significance exists in the data.

In the previous example of comparing the average gas mileage of two makes of automobiles, the decision maker could decide that a difference of one mile per gallon or more is operationally significant; any difference less than one mile per gallon is not. Suppose, also, that the decision maker is willing to use a significance level of 5 per cent and a Type II error probability of 10 per cent. The statistician can now specify how many Fords and Plymouths should be tested to detect the operationally significant difference of one mile per gallon in average gas mileage of each make of automobile at the 5 per cent level of statistical significance.* If less than the specified number of automobiles can be made available for testing, the experiment should not be performed, i.e., the data from the experiment will be insufficient to answer the question of a difference existing of one mile per gallon or more in average gas mileage between the two makes of automobiles.

Test Statistic

Fixing the significance level for a particular type of statistical test on experimental data determines a fixed numerical quantity, called a critical value, with which the statistician can compare the

---

* This assumes that the variability of gas mileage between members of each automobile population is known.

test statistic. The test statistic is a numerical quantity which the statistician computes from the experimental data and, therefore, depends on the numerical values of the data. By comparing the relative magnitude of the test statistic with the critical value, the statistician can determine whether or not the null hypothesis of the experiment should be accepted or rejected at the given significance level. Essentially, this is how the statistician performs hypothesis testing.

Replication

Replication is the repetition of an experiment, under as identical conditions as possible, on a different experimental unit or subject. Repeated measurements on the same experimental unit or subject are not usually considered as replications.

Normally, field experiments are conducted to gain information about characteristics of some type of large population. It is desirable to be able to infer that results obtained from testing a sample of a population also apply to the population itself. The larger the size of the sample or the greater the number of replications in an experiment, the more confidence is gained that results from the experiment are applicable to the population from which the sample was drawn. Increasing the sample size provides more assurance that representative units of the population are included in the sample.

The number of replications in an experiment should be synonomous to the sample size of the experiment. In the ammunition

example, suppose only one round was fired and its muzzle velocity measured. Could one be very sure that the muzzle velocity of that round is truly representative of what could be expected from the whole ammunition lot? Repeated measurement on the same sample unit obviously is not possible. Additional rounds must be fired or replicated to increase confidence in the experiment's results. Absolute assurance about the acceptability of the lot could be gained if all the ammunition were fired, but then no ammunition is left to accept. Thus, an economically large enough sample should be tested to be reasonably confident that the results of the experiment are applicable to the whole ammunition lot. In this case, the number of replications clearly would be equal to the sample size.

Repeated observations or measurements on the same sample unit serves to provide a more precise estimate of the true value of the characteristic being measured for that sample unit only. That particular sample unit still may not be representative of the overall population. Which of the following methods would be preferable for estimating the average gas mileage of the overall population of Ford automobiles? Test each of ten Fords once and use the sample mean gas mileage of the ten as the estimate; or, take one Ford and test it ten times and use the sample mean of the ten tests on one Ford as the estimate. The latter method certainly provides a very precise indication of what the actual average gas mileage is of the one automobile, but that method provides little assurance that all

Fords have that particular gas mileage on the average. Definitely the former method is preferable to meet the objective of the experiment, estimating the average gas mileage of the Ford population.

In the example of the comparison of average gas mileages of two makes of automobiles, knowledge of the sample mean gas mileage of each make of automibile is by itself insufficient to permit hypothesis testing. Knowledge of the variability of the gas mileage of each sample unit tested about its respective sample mean is also important. Variability is caused by actual differences between sample units and by experimental error in the selection and testing of the sample units. This variability can be referred to as the sample variation or sample standard deviation.* Replication produces experimental data from which the computation of the sample variance is possible. Since the sample variance is a required component of the computational formulae of the test statistic used in hypothesis testing, hypothesis testing is not possible without replication.** Thus, without adequate replication in the experiment, there is little, if anything, that could be said about statistical differences between the average gas mileage of the two makes of automobiles.

---

* Sample standard deviation is the positive square root of sample variance.

** At least 2 replications are necessary if the variability of data is unknown prior to the conduct of an experiment in order to use the hypothesis testing technique.

The number of replications in an experiment are adequate if, for a given significance level, the probability of detecting a specified operationally significant difference, if it actually exists, is high. In general, for a given significance level, the smaller or more refined that an operationally significant difference becomes, the more replications are needed to detect that difference at the same level of Type II error probability. In the example, fewer automobiles would need to be replicated to detect a difference of five mpg than would need to be replicated to detect a difference of one mpg while maintaining the same level of Type II error probability for both sample sizes.

The Relationships Among Type I Error, Type II Error, and Replication

The relationship between Type I and II error has been discussed on page 31 of this chapter. Recall that for a fixed number of replications Type II error probability could be decreased at the expense of an increase in Type I error probability. The only way to decrease both Type I and Type II error probabilities simultaneously is to increase the number of replications in the experiment.

Normal procedure in designing simple experiments, where the expected variability of data is essentially known, is to fix what are acceptable levels of Type I and Type II error probabilities, determine what is operationally significant, and then determine the number of replications necessary to be performed. Since this

procedure is used in a large variety of industrial experiments, especially in quality control work, standard figures or graphs have been developed from which the statistician can quickly determine the necessary number of replications to meet specified probabilities of Type I and II errors. These graphs are called operating characteristic curves. The curves prescribe the operating characteristics, Type I and II error probabilities, number of replications, and operational significance, for a hypothesis test.

Table IV reflects the interaction between the number of replications, and Type I and II error probabilities which occur for certain experimental results.[4] Notice that for a fixed Type I error probability, the Type II error probability decreases as the number of replications increases, e.g., for a Type I error probability of 5%, the Type II error probability of 92% at 2 replications is decreased to 2% for 20 replications. Also notice that for a fixed Type II error probability, the Type I error probability decreases as the number of replications increases, e.g., for a Type II error probability of 20%, the Type I error probability of 5% at 10 replications is decreased to a Type I error probability of 1% at 15 replications.

The strength or power of a statistical test is often called the power of the test. The power of a test is also a probability, one

---

[4] Ibid. p. 32, d=1 for a variance and operational significance of one unit.

TABLE IV

RELATION BETWEEN REPLICATIONS
AND TYPE I AND II ERRORS*

Number of Replications	Probability of a Type I Error		Probability of a Type II Error	
2	5%	(1%)	92%	(98%)
3	5%	(1%)	83%	(95%)
4	5%	(1%)	72%	(92%)
5	5%	(1%)	62%	(87%)
7	5%	(1%)	41%	(72%)
10	5%	(1%)	20%	(50%)
15	5%	(1%)	5%	(20%)
20	5%	(1%)	2%	(7%)

*Type I and II errors probabilities are related by the omission or inclusion of parentheses, e. g., 2 replications and a Type I error probability of 5% produces a Type II error probability of 92%.

minus the probability of a Type II error, e.g., if Type II error probability is 20%, the power of the test is 80%. The power of a test is the probability of rejecting the experiment's null hypothesis when, in fact, the null hypothesis is false. Since the aim of an experiment is to attempt to reject the null hypothesis of the experiment when it is false, it is desirable to be able to use a powerful or strong statistical test.

It has been shown that for fixed Type I error probability, the Type II error probability decreases as the number of replications increases. As Type II error probability decreases, the power of a test increases. Therefore, by increasing the number of replications, the power of a statistical test can be increased, i.e., the probability of detecting a truly operationally significant difference is increased.

Confidence Intervals

Confidence intervals become important when the statistical problem faced is one of estimation. The purpose of an experiment to answer an estimation problem is to attempt to describe a characteristic of a population. For example, the problem might be to estimate the average gas mileage of 1968 Fords with no attendant need to perform hypothesis testing. "What is the average gas mileage?" is the question. The sample mean of the experimental data could be used as the estimate of the average gas mileage.

A point estimate which describes a population's characteristic with a specific numerical value should always be presented with a

confidence interval. Failure to use a confidence interval about an estimate lends a false sense of exactness to that estimate. Suppose that a sample of the 1968 Ford population is tested to determine the average gas mileage of Fords. Consider the following imaginary conversation between a Ford Motor Company executive, Henry, and his statistician, Dr. X.

Henry - "Well, Dr. X, what can we tell our potential customers about the gas mileage to expect from our 1968 Fords with the Firebelch engine?"

Dr. X - "Sir, our recent experiment indicates that the average gas mileage is 16 per gallon over the type of driving conditions tested."

Henry - "Are you sure that 16 mpg is correct?"

Dr. X - "No Sir, but I am 95 per cent confident that that the mean gas mileage is 16 mpg plus or minus 2 mpg."

Henry - "Why do you say plus or minus 2 mpg?"

Dr. X - "I am not sure that the exact mean gas mileage is 16 mpg; but if I am permitted to be in error by plus or minus 2 mpg, I can be 95 per cent confident that the true mean gas mileage is between 14 and 18 miles per gallon."

Henry - "Can't you be more confident than 95 per cent?"

Dr. X - "Yes Sir, I can if I increase my tolerance for error. For example, I am 99 per cent confident that the mean gas mileage is 16 mpg plus or minus 10 mpg."

Henry - "That's not good enough for our advertising campaign. I want our Madison Avenue boys to be able to say the Firebelch engine in our automobile will produce an estimated average gas

mileage plus or minus one-half mpg with a confidence of 99 per cent. What do you need to do?"

Dr. X - "I need to substantially increase the number of automobiles tested from 10 to 85. Shall I test an additional 75 automobiles?"

Henry - "The Comptroller will have a fit, but go ahead. You know how stringent the regulations on truth in advertising are getting."

Dr. X could have been 100 per cent confident about the exact value of the average gas mileage if he had accurately tested the entire population of Fords in question; or, he could have been 100 per cent confident about the sample results if he made his confidence interval large enough, say 0 mpg to 100 mpg.

The important points to note in the example are that for a given sample size, the degree of confidence associated with a confidence interval can be increased or decreased by a respective increase or decrease in the width of the interval; and, if the degree of confidence associated with an interval is specified, the size of the interval is decreased only by increasing experimental replications. Table V illustrates that a relationship similar to that which exists among Type I and II error and replication also exists among the degree of confidence, size of the confidence interval, and the number of replications.

Once an experiment has been conducted and a population's characteristic has been estimated and a specific confidence interval determined, it is not proper, in a probabalistic sense, to say that

TABLE V

RELATIONSHIP AMONG DEGREE OF CONFIDENCE, SIZE OF CONFIDENCE INTERVAL, AND NUMBER OF REPLICATIONS

Degree of Confidence	Size of Confidence Interval	Number of Replications Required
Increases	Increases	Fixed
Decreases	Decreases	Fixed
Increases	Fixed	Increases
Decreases	Fixed	Decreases
Fixed	Decreases	Increases
Fixed	Increases	Decreases

there is a certain probability that the specific confidence interval includes the actual value of the characteristic being estimated. To avoid this difficulty, statisticians employ the notion of a degree of confidence about whether or not the actual value is included in a specified interval. Either the actual value is or is not included in any specific confidence interval calculated from experimental data. The notion of 95 per cent confidence, for example, comes from a feeling that if the experiment were exactly repeated 100 times on 100 different samples, the statistician would expect 95 of the confidence intervals generated from the 100 sets of data to include the actual value of the characteristic being estimated.

Suggested Readings

For the reader who is interested in further pursuance of an understanding of statistics, a list of recommended references is presented. This list is by no means inclusive of the number of good books available on statistics, but it will serve as a starting point for a person who desires to read more about statistics.

The first four references are very non-technical and virtually assume that the reader knows nothing about statistics when he opens the cover of the book. The remainder of the references are of the textbook type and require a knowledge of simple calculus and basic probability theory to follow some of the mathematical derivations contained in them. However, the non-mathematician should not be discouraged by seemingly awesome symbols, formulae, and equations

in these references. A great deal of insight to statistics can be gained even if the mathematical portions of the texts are ignored. For example, Chapter 6 of the fifth reference provides an easy-to-read introduction to statistical decision theory which requires little more than a knowledge of arithmetic and simple algebra.

1. M. J. Moroney, "Facts From Figures", Penguin Books, Baltimore, 1964.

2. W. J. Reichmann, "Use and Abuse of Statistics", Oxford University Press, New York, 1962.

3. A. N. Franzblau, "A Primer of Statistics for Non-Statisticians", Harcourt, Brace & Co., New York, 1958.

4. D. Huff and I. Geis, "How to Lie with Statistics", Victor Collancz Limited, London, 1954.

5. S. Ehrenfeld and S. B. Littauer, "Introduction to Statistical Method", McGraw - Hill, New York, 1964.

6. H. Chernoff and L. E. Moses, "Elementary Decision Theory", John Wiley & Sons, New York, 1959.

7. B. Ostle, "Statistics in Research", Iowa State University Press, Ames, Ia, 1964.

8. R. Goodman, "Modern Statistics", Arc Books, New York, 1964.

9. A. H. Bowker and G. J. Lieberman, "Engineering Statistics", Prentice Hall, Englewood Cliffs, N. J., 1959.

10. C. R. Hicks, "Fundamental Concepts in the Design of Experiments", Holt, Rinehart, & Winston, New York, 1966.

11. W. S. Ray, "An Introduction to Experimental Design", MacMillan Co., New York, 1960.

12. R. A. Fisher, "The Design of Experiments", Hafner-Publishing Co., New York, 1947.

13. N. L. Johnson and F. C. Leone, "Statistics and Experimental Design in Engineering and the Physical Sciences", Vol. I, John Wiley & Sons, New York, 1964.

14. A. E. Mace, "Sample-Size Determination", Reinhold Publishing Corp., New York, 1964.

Each of the above references will themselves contain extensive listings of bibliographies and suggested reading references. All of the references listed above are available at either the reference library or the text issue facility of the Naval Postgraduate School.

CHAPTER IV

DISCUSSION OF AN EXPERIMENT

This Chapter discusses an experiment conducted by the United States Army Combat Developments Command Experimentation Center in April of 1964. Primary emphasis will be given to the application of statistical techniques used in the experiment, especially the application of the statistical concepts discussed in Chapter III. However, since statistical techniques are applied to data which quantify a measure of performance, it is also necessary to understand the measures of performance used in the experiment if the inferences drawn from the statistical interpretation of the data are to be operationally meaningful. A critique of the application of statistical techniques and measures of performance used in the experiment will be made as the experiment is described and its results discussed. The Chapter concludes with a summary of comments about the experiment.

The quality of the report of the experiment should not be considered as typical of the experimental reporting done by the Experimentation Center. This particular experiment was purposely selected because it is felt that certain weaknesses exist in the design of the experiment and in the analysis and findings of the experiment as presented in the experiment's report.

Other considerations for using the experiment as an example

were: The experiment's report is unclassified;* the objective and scope of the experiment were limited and can manageably be discussed in this chapter; the purpose of the experiment is representative of the Army's need to perform field experimentation; the report has been approved for distribution through the Defense Documentation Center.

I. THE EXPERIMENT

The experimental report is "Comparison of Fire Effectiveness - Mounted vs Dismounted", USACDCEC, June 1964. The description of the experiment is summarized from the experiment's report.

The purpose of the experiment was to provide input information for a Department of Army comparative evaluation of armored infantry doctrines.

The experiment's objective was "to compare the effectiveness of fire by troops from moving tracked vehicles and fire by dismounted troops."[5]

Measures of Fire Effectiveness

"Ability to defeat a point target" and "ability to place suppressive fire in an area" were the criteria used to determine fire effectiveness. The measures of performance used to describe "ability to defeat a point target" were accuracy and time to obtain

*"For Official Use Only" restrictions were removed January 7, 1967 under authority of Army Regulation 345-15.

[5]Section II, p. 3.

a first hit. The measures of performance used to describe "ability to place suppressive fire in an area" were an index of target area hits and the volume of effective fire. The type of data produced by the experiment to quantify each measure of performance will be defined later in the chapter when the experimental results are discussed.

The report is unclear as to exactly what type of combat element is being examined by the experiment. Whether the combat element to be tested is a mounted or dismounted individual or a mounted or dismounted unit is not specified. Although this is a question of problem definition, it is pertinent to the statistical design of the experiment. Should experimental replication be performed on individual firers or on units composed of individual firers? The type of sample unit to be tested in the experiment should have been clearly defined in the report.

Resources

Time was apparently a critical factor; at most, only eight days were spent in the field conducting trials and collecting data. A total of twenty-four riflemen and twelve machinegunners armed with M14 rifles and M60 machineguns were subjects. M113 and M106 armored personnel carriers were modified to permit mounted personnel to fire forward as the vehicles moved forward over two different firing courses. No monetary constraints are mentioned in the report.

Firing Courses and Traverse Speeds

One firing course was over "smooth" terrain; the other was over "rough" terrain. Each of the two courses had 3 groups of targets in its target area. The dismounted troops traversed both courses at the same speed (average 2.25 mph). The mounted troops traversed both courses at two different speeds, "slow" (average 7.5 mph) and "medium" (average 11.2 mph). On each course, firing commenced 200 meters from the target area and ceased fifty meters from the target area; i.e., troops traversed 150 meters while firing. No rationale was given in the report as to why 200 meters was the commence fire line.

Target Arrays

Each of the three target groups per course had seventy-seven targets. One target in each group was painted white. The white target was centrally located in the target group and was the aiming point for troop firings. The remaining seventy-six targets in the group were designated as sensor targets to sample the impact of rounds in the target area. All targets were "E" type silhouette targets. The white targets were remote controlled pop-up targets; other targets were stationary.

The dimensions of the target area of each course were fifty to seventy meters deep and approximately fifty-five meters wide.

The reader should keep in mind that the target arrays constructed for this experiment do not represent a typical enemy defensive threat

posture. For this reason, it may be difficult to relate the results of the experiment to actual combat situations.

Fire Discipline

Dismounted riflemen paused every few steps to fire semi-automatically from the shoulder. Mounted riflemen fired automatically from the shoulder. Dismounted machinegunners paused every few steps to fire short bursts from an underarm position. The mounted machinegunners fired their guns, fixed in flexible mounts on the carriers, in short bursts from the shoulder. All firers were directed to attempt to strike the white target.

During the conduct of a trial, all troops would fire at one white target while traversing fifty-meter increments of the course. After traversing fifty meters, the white target in a new target group would appear and fire would be transferred to the new group, the white target as the aiming point. Each of the three target groups on a course came under fire once during a trial. When raised, a white target remained up throughout a fifty meter traverse period, i.e., a hit on the white target did not depress or "kill" the target.

Subject Personnel

The twenty-four riflemen were organized into two 6-man groups for the mounted and dismounted modes. The twelve machinegunners were organized into two 3-man groups for each mode. The weapons qualification scores of the troops were used as a basis to assign

individual firers to a group such that each group had roughly the same average (mean) score per man.

The report states: "For each weapon type [M14 or M60], 10 per cent [of the subjects] were Experts, 40 per cent were Sharpshooters or first class gunners, and 50 per cent were Marksmen or second class gunners."[6] The statement is footnoted in the report - "This is the standard breakdown of firing qualifications throughout the Army." Note that 10 per cent of twenty-four riflemen is 2.4 Experts, 40 per cent is 9.6 Sharpshooters, and 50 per cent is 12 Marksmen.

Certainly it is ridiculous to think that each group of riflemen was composed of 0.6 Experts, 2.4 Sharpshooters and 3 Marksmen. But, the wording of the report implies that each six-man group is relatively the same and is representative of the Army in its group marksmanship ability. To be close to the stated percentages, either two (8.3%) or three (12.5%) riflemen were Experts, and either one (8.3%) or two (16.7%) machinegunners were Experts. Although mean scores were similar for each group, actual composition of the groups must have been different in the number of Experts, Sharpshooters, and Marksmen assigned to some of the groups.

For this experiment, it would have been preferable to equate individual shooting ability between groups rather than attempt to meet the 10-40-50 per cent breakdown on all firers, e.g., each

---

[6]Annex A, Para 2f, p. 19

rifleman group could have been composed of one Expert, two Sharpshooters, and three Marksmen. Differences between groups could then be reasonably attributed to the mounted or dismounted mode and not influenced by group composition.

Replication of Trials

Once personnel were segregated into mounted and dismounted groups, they remained as mounted or dismounted subjects throughout the experiment.

The report states: "To insure that data were statistically valid, thirty-six repetitions of each [terrain] condition for each weapon mode were performed by individual riflemen; eighteen repetitions of each condition were performed by individual machinegunners. This resulted in a grand total of 324 repetitions."[7] An inconsistency is now apparent. Trials were conducted on groups of six riflemen and on groups of three machinegunners. Replication of the trials appears to be by groups since data was recorded on the basis of a group firing. No data was recorded for individual subjects; individual performances were merged and recorded as a group performance. Hence, it is rather meaningless to note individual repetition since individuals were firing as a group. Actually, only two replications of the experiment were conducted on each firing course since subjects were structured into two groups per weapon- mode and subjects remained

[7]Ibid. 5. p. 27

in the mode to which they were originally assigned.

The report is very unclear about the structure of group trials. Apparently, each dismounted group was put through each terrain course three times; and each mounted group was put through each terrain course three times at slow speed and three times at medium speed. This is repeated measurement on a group and not experimental replication. Machinegunner and rifleman trials must have been separate since rifle and machinegun results are separated in the report.

At least implicity, the previous question of what combat element is being studied has been answered.* The experiment actually measures the "abilities" of mounted groups vs dismounted groups. The sample unit of this experiment is the mounted or dismounted group.

II. FINDINGS OF THE EXPERIMENT

The analysis of data generated to meet the objective of this experiment is particularly amenable to hypothesis testing. Yet, no null hypothesis was formulated and no hypothesis testing performed.

Consider now what results were reported and the critiques of those results. Table VI is a summary of data compiled and condensed from Annex B of the experiment's report.

---

*Page 50.

TABLE VI

SUMMARY OF TRIAL DATA

Weapon	Terrain	Mode*	1 Hits on Pop Up Targets	2 Hits on Sensor Targets	3 Rounds Expended	Mean Traverse Time (Sec)	Hit Probability (1/3)	Index of Target Area Hits (2/3)
M14 Rifle	Smooth	DM	164	1637	3380	181.5	.048	.484
	Smooth	MSS	81	1349	3275	47.5	.025	.411
	Smooth	MMS	55	1025	2503	27.7	.022	.410
	Rough	DM	95	654	3309	148.6	.029	.198
	Rough	MSS	31	448	2962	42.3	.010	.151
	Rough	MMS	23	378	2355	28.2	.010	.161
M60 Machine Gun	Smooth	DM	113	2266	4493	139.8	.025	.505
	Smooth	MSS	146	2671	4400	42.7	.033	.607
	Smooth	MMS	160	2547	4117	28.8	.039	.618
	Rough	DM	44	882	4494	123.8	.010	.196
	Rough	MSS	56	1041	4242	48.5	.013	.246
	Rough	MMS	36	647	3901	32.8	.009	.161

*DM - Dismounted, MSS - MOUNTED Slow Speed, MMS - MOUNTED Medium Speed

Accuracy

The accuracy measure of performance for a weapon-mode was quantified as the ratio of total hits on the white pop-up targets to total rounds fired, i.e., the hit probability column of Table VI. Hypothesis testing could have been used very effectively on this measure had sufficient replication been performed. The null hypothesis could have stated "the single round hit probability of dismounted groups, mounted groups at slow speed, and mounted groups at medium speeds are equal."

The findings of the report are stated as follows:

> Dismounted M14 riflemen were two [on smooth terrain] to three [on rough terrain] times as accurate [higher hit probability] as mounted riflemen at either speed. . . .[8]

and,

> In the smooth terrain the mounted M60 machinegunners were more accurate at both speeds then were the dismounted gunners, but in the rough terrain accuracy at the medium speed was no more than that when dismounted.[9]

Note that the findings of the report are drawn from a comparison of magnitudes of point estimates. Recall that a point estimate should always have a confidence interval describing a degree of belief in the amount of error involved in making the estimate. Since only two replications of each weapon-mode was performed, the confidence interval width for the estimates of hit probability would reasonably

---

[8]Section III, para 1a(1), p. 4

[9]Ibid. 1a(2).

be expected to be relatively large for any high degree of confidence.* Suppose that a 95 per cent confidence interval required an error of plus or minus .020 about the hit probability estimates for riflemen on smooth terrain. The resulting 95 per cent confidence interval for dismounted is from .028 to .068; for mounted-slow it is from .005 to .045; for mounted-medium it is from .002 to .042. Because of the margin for error in the confidence interval, it is possible that the true hit probability of all three riflemen modes could be .030, i.e., there is no true difference, and a Type I error has been committed by concluding that one mode is more accurate than another.

Although the mathematical structure of hypothesis testing does not involve a direct comparison of confidence intervals, hypothesis testing is a procedure which takes into account the margins of error about a point estimate in determining whether or not the null hypothesis should be rejected. If only one group had been replicated in each mode, no hypothesis testing could have been performed.** With two replications in each mode, hypothesis testing would have been possible, but the margin of error could be so great as to result in no determination of a significant difference, i.e., a failure to reject the null hypothesis. A failure to reject the null hypothesis allows the possibility of the commission of a Type II error. Because only

---

*Depending on the amount of variability in the data.

**Assuming that the variability of data was not known prior to the conduct of the experiment.

two replications were performed, it is reasonable to assert that the probability of a Type II error in a hypothesis test on the data of this experiment would be very high for any low Type I error probability. A meaningful hypothesis test with low probabilities of Type I and Type II error could be performed only if the experiment had included a greater number of replications on each of the mounted and dismounted modes, i.e., if the sample size had been enlarged for each mode.

The report could have included data on the performance of each of the two groups replicated in a particular mode. This would have provided some indication of the variability of the data. Suppose, on the smooth terrain, that the two dismounted riflemen groups had hit probabilities of .028 and .068, the average of which is .048, and that the two mounted-slow riflemen groups had hit probabilities of .035 and .015, the average of which is .025. Can one now say that dismounted riflemen are almost twice as accurate as mounted riflemen? A group-by-group comparison of hit probabilities requires a negative answer to this question. At least one would feel a little better about the report's findings if it could have been reported that the individual group hit probabilities did not vary "too far", say within .002, of the average of both groups.

Consider the findings on mounted and dismounted machinegunner accuracy. Since the usual employment of machinegun fire in an attack is to achieve a suppressive or "covering" effect on the enemy,

it is questionable whether or not it is operationally meaningful to attempt to compare the accuracy of mounted and dismounted machinegun fire against a point target. Therefore, it is questionable whether or not any difference in hit probabilities between mounted and dismounted machinegunners is operationally significant.

But, assuming that a difference between machinegunner's hit probabilities is operationally meaningful, by omission of a contrary statement, the findings of the report infer that the mounted machinegunners were more accurate than the dismounted machinegunners on rough terrain. Even without hypothesis testing, is a difference in the hit probability estimates between .010 and .013 really operationally significant? The change in magnitude is very small. Does it seem reasonable that doctrinal and tactical changes should be made, probably at great expense, to achieve a .003 addition to a machinegunner's accuracy or hit probability? Granted, an increase from .010 to .013 is a 30 per cent increase. However, the increase really means that for every 400 rounds fired by a machinegunner, the mounted machinegunner could expect 5.2 hits and the dismounted machinegunner could expect 4.0 hits. Even if an addition of .003 to the hit probability for machinegun fire is defined as operationally significant, it is doubtful that a hypothesis test analysis of data produced by this experiment would have detected any statistical significance since only two replications were performed.

What would appear to be operationally significant and meaningful

is that the experiment's results indicate that both dismounted and mounted machinegunners were at least as accurate as mounted riflemen on both terrain courses. This lends evidence to infer that all mounted firers should be armed with machineguns if accuracy of fire is desired.

To draw conclusions from an "eyeball" comparison of the magnitude of estimates is to ignore the existence of Type I and II errors which are inherent to experimentation. Even if several groups of mounted and dismounted riflemen and machinegunners had been replicated, formal hypothesis testing is much preferred to the "eyeball" test if the conclusions inferred from the analysis of data are to be statistically sound. The "eyeball" test may have its place in troop testing, but it should definitely be avoided in field experimentation.

The reader should also be aware that to use a ratio of hits to rounds fired and to call that ratio a single-round hit probability requires an assumption of independence of each round fired.* In this case, independence means that each round fired has an equally likely chance of striking the target. The assumption seems reasonable when applied to dismounted riflemen firing with a semi-automatic cycle. However, one should seriously question the assumption if it is also applied to weapons firing with an automatic cycle of fire, especially on dismounted machinegunners who may have a tendency

---

* This assumption is noted in the report, p. 4.

to "walk" fire up to a target. Independence of rounds implies that there is no learning effect in aiming the rifle or machinegun by observing the impact of rounds previously fired.

Time to Obtain a First Hit

"Time to obtain a first hit" is actually a misnomer of the measure of performance which is analyzed in the report. One might feel intuitively that it would measure the length of time from the instant when group fire commenced on a white target until the white target was hit for the first time. It is an unfortunate choice of words. What is discussed in the report under the title of "Time to Obtain a First Hit" is the expected number of hits achieved by an individual firer during a nine, ten, and fifteen second time interval.* The report explains the measure of performance in the following manner.

> These periods of 9, 10 and 15 seconds during which mounted firers traversed range segments are of suitable lengths for use in evaluating a firer's reaction capability when confronted with a target; that is, his ability to achieve hits in these time periods indicates in a general way how much of a chance the firer has of achieving a hit before the enemy hits him. . . . .[10]

How, then, was the data analyzed to draw a comparison between mounted versus dismounted firers?

Recall that a different white target was exposed and fired upon

---

[10] Ibid. 1b(1)(b), p. 6

*Perhaps "hit rate", the number of hits per unit of time, would have been a more appropriate title for the analysis of this data.

during each fifty meters of course traverse during a trial. Table VII presents the average traverse times over a fifty-meter segment per individual firer.* The averages are computed from data captured on both courses. For example, the mean traverse time of dismounted riflemen on the smooth and rough course in Table VI are added, producing 330 seconds. Dividing 330 seconds by six (three 50-meter segments per course) gives the result of fifty-five seconds in Table VII. Similarily, the total hits, 164 and 95 from Table VI, on each course are added to a total of 259 hits for dismounted riflemen. The 259 hits is divided by 216 to give the result of 1.2 hits in Table VII. Where does the 216 come from? Two dismounted groups of six men traversed each course of three 50-meter segments three times: 2 x 6 x 2 x 3 x 3 = 216.

TABLE VII

AVERAGE TRAVERSE TIME AND HITS PER 50-METER SEGMENT PER INDIVIDUAL FIRER

	Riflemen		Machinegunners	
Mode	Time (sec)	Hits	Time (sec)	Hits
Dismounted	55	1.2	44	1.5
Mounted Slow	15	0.5	15	1.9
Mounted Medium	9	0.35	10	1.8

*
The data is the same as in Table I, p. 6, of the report.

The report continues:

> In order to compare the abilities of dismounted firers with those of mounted firers it is necessary to calculate the hits that a dismounted rifleman could expect to achieve in 9 and 15 seconds and a dismounted machine-gunner in 10 and 15 seconds. The procedure used for the rifleman calculations was as follows:
>
> It was assumed that riflemen's hits during a 55-second average traverse time were distributed almost equally throughout that period. . . . . Average hits during the 55-second period were then converted mathematically to expected average hits during the shorter periods."[11]

The "mathematical" conversions of dismounted riflemen hits to the shorter periods are: 1.2 hits is divided by 55 seconds the result of which is multiplied by 15 seconds and 9 seconds producing 0.33 hits in 15 seconds and 0.20 hits in 9 seconds, respectively. By the same method of conversion, dismounted machinegunners could expect 0.5 hits in 15 seconds and 0.3 hits in 10 seconds.

A comparison of the results of the above paragraph with entries in Table VII reveals that the expected number of hits for dismounted firers are less than mounted firers for each respective time period of 9, 10, and 15 seconds. The report concludes that "Mounted riflemen and mounted machinegunners achieve hits substantially more quickly than do dismounted firers."[12]

The criticism of comparing point estimates also applies to the report's analysis of this measure of performance. Again the

---

[11] Ibid. 1b(1)(c), p. 6

[12] Ibid, IV, - a, p. 15

questions about confidence intervals, hypothesis testing and Type I and II errors are pertinent. The information in Table VII is not separated for the two terrain courses or by mounted or dismounted groups; information is combined into an overall average per individual when, actually, groups of firers were tested in each trial. The statistician would be hard pressed to determine any confidence intervals or to perform hypothesis testing on Table VII information since the value of replication is lost when averages of averages are averaged for the whole experiment; there is no way to determine the variability of the information in Table VII.

Does the report's analysis really answer the question of whether a mounted or dismounted firer has a better chance of hitting the enemy before he hits him, even in a "general way"? It appears that the analysis contains what statisticians sometimes refer to as Type III error, an answer to the wrong question. Of course dismounted firers have an average of fewer hits per unit of time. They fired fewer rounds per unit of time. Table VIII contains information on the average rate of fire of mounted and dismounted troops which was not included in the report.*

Data which could have answered the question, i.e., the measurement of the elapsed time between presentation of a target of opportunity

*Average individual rates of fire in Table VIII was computed from the data contained in Table VI; i.e., rounds expended divided by the product of the number of individual repetitions and mean traverse times.

TABLE VIII

AVERAGE RATE OF FIRE
OF INDIVIDUAL FIRERS

Weapon	Terrain	Mode	Rate of Fire (rounds per second)
M14*	Smooth	Dismounted	0.52
	Smooth	Mounted Slow	1.92
	Smooth	Mounted Medium	2.54
	Rough	Dismounted	0.62
	Rough	Mounted Slow	1.94
	Rough	Mounted Medium	2.43
M60	Smooth	Dismounted	1.78
	Smooth	Mounted Slow	5.72
	Smooth	Mounted Medium	7.95
	Rough	Dismounted	2.02
	Rough	Mounted Slow	4.95
	Rough	Mounted Medium	7.28

*Recall that dismounted riflemen fired on a semi-automatic cycle and mounted riflemen fired on a full automatic cycle.

and the target receiving the first hit, was not captured by the experiment. Only with such data could a meaningful analysis be performed between the more accurate dismounted riflemen and the faster firing but less accurate mounted riflemen. What is more important in achieving a first hit? Is it accuracy of fire or rate of fire or a certain combination of both? The analysis of the report does not consider the relationship of these important causal effects. Proper measurement of time to first hit is necessary to provide a data base in which accuracy and rates of fire are interrelated in the data. A controller with a stop watch and a "killable" white target* for each target group would have been the only additional resources needed to measure time to a first hit in this experiment.

Index of Target Area Hits

This index measured the "ability to place suppressive fire in an area" and is tabulated in the right most column of Table VI on page 56. The index is computed by dividing the total hits on sensor targets by the number of rounds expended by all firers per weapon, mode, and terrain course.

According to the report, the higher the index, the better is the ability to place suppressive fire in an area. The findings in the report are stated as:

*Trainfire type "kill" devices for pop-up targets were in common use in 1964, especially at basic training centers such as Ft. Ord where USACDCEC is located.

> In both terrains the dismounted riflemen hit sensor targets with a higher percentage of their rounds fired than did the mounted riflemen at either speed; the mounted riflemen performed equally well at the two speeds.[13]

and:

> In both terrains the mounted machinegunners at slow speed hit sensor targets with a higher percentage of their rounds fired than did the dismounted gunners. There was virtually no difference between mounted gunner performance at the two speeds in the smooth terrain, but mounted performance at the medium speed was poorer than either of the other conditions in the rough terrain.[14]

Even if the "eyeball" comparison of point estimates was a proper statistical technique and even if the above conclusions were restricted to apply solely to the firers tested in the experiment, a degree of uncertainty should exist about the results for this measure of performance. The report acknowledges that sensor targets only sampled impacts in the target area. All rounds fired into the target area were not recorded as hits on sensor targets. Hence, the results require a very strong assumption, not stated in the report, that the proportion of sensor target hits to total rounds actually landing in the target areas is equal among dismounted firers and mounted firers at both speeds. Only with this assumption could one even begin to infer that a higher index of target area hits indicates better performance. Furthermore, only by using the technique of hypothesis

---

[13] Ibid. III, 2a(2), p. 9

[14] Ibid., 2a(3), p. 10

testing and a subsequent rejection of the null hypothesis would the inference of a true difference be statistically valid.

Again, since only two replications of dismounted groups and mounted groups at both speeds were conducted in the experiment, it is entirely possible that a hypothesis test would not reveal any differences in the index measure of performance among comparable groups. Also, the probability of a Type II error would be very high for any selected low significance level, say below 20 per cent.

Volume of Effective Fire

The report defines the volume of effective fire as follows:

> This measure is defined as the average number of rounds of an individual rifleman or machinegunner hitting the sensor targets per second. These rounds strike the target area at essentially random locations; thus this volume of fire is a direct indicator of suppressive effect at the enemy position, because it will determine how often, and for how long, the defending enemy will feel compelled to "pull their heads down".[15]

Table IX presents the numerical values for this measure of performance.* The computational method to determine average hits per firer per second is to divide the hits on sensor targets by the product of mean traverse time and the number of individual repetitions with each weapon, 36 for riflemen and 18 for machine-gunners.** An alternate method of computation which gives

---

[15] Ibid. 2c(1), p. 11.

* The information was presented in bar chart form in the report.

**Hits and traverse time data are found in Table III, p. 30.

TABLE IX

AVERAGE HITS ON SENSOR TARGETS
PER SECOND OF INDIVIDUAL FIRERS

Weapon	Terrain	Mode	Average Hits Per Second
M14	Smooth	Dismounted	0.25
	Smooth	Mounted Slow	0.79
	Smooth	Mounted Medium	1.03
	Rough	Dismounted	0.12
	Rough	Mounted Slow	0.29
	Rough	Mounted Medium	0.37
M60	Smooth	Dismounted	0.90
	Smooth	Mounted Slow	3.47
	Smooth	Mounted Medium	5.02
	Rough	Dismounted	0.40
	Rough	Mounted Slow	1.20
	Rough	Mounted Medium	1.10

approximately the same results is to multiply the index of target area hits of Table VI, page 56, by the average rate of fire of Table VIII, page 66, for each respective weapon-terrain-mode.

True to form, the report finds that "both riflemen and machine-gunners placed at least twice as much fire per second into the target area when mounted as when dismounted."[16]

The critique of the use of this measure of performance and of the conclusions drawn is left to the reader with the following questions, not answered by the report, as a guideline.

1. Why is the assumption of hit proportionality on sensor targets discussed on page 68 implicit to this measure of performance?

2. Recall that each target array had three target groups. The two-dimensional profile facing the firer of each target group was 15 meters wide and 10 meters high. Is it reasonable to assume that each round passing through this profile "window" would have the same suppressive effect on the white target which simulated the enemy? For example, the analysis implicitly assumes that rounds striking sensor targets 5 meters to the right or left of the white target would have an equal suppressive effect as rounds striking sensor targets 1 meter to the right or left of the white target.

3. What is the necessary average number of hits per second "near" an enemy to keep his head down and why might this be important in the analysis?

4. Suppose that one round every two seconds striking near an enemy is necessary to suppress him. Since the calculation of average hits per second was based on a sample of the total rounds actually striking the area of a target group, what conclusions can be drawn by applying the report's analysis?

---

16 Ibid. 2c(2), p. 11

Suppose, further, that an assumption is made that the sensor targets sampled about one-half of the rounds striking the area of a target group.* Would not this imply that a dismounted rifleman on smooth terrain can continuously suppress a point target, which is the desired effect being measured? Then, would not a firer with average hits per second in excess of 0.25, based on the sensor target hits sample, just be "over-suppressing" the target enemy?

5. Why would an analysis using hypothesis testing on the average hits per second per group of firers have been more meaningful, especially if sensor target hits were "weighted" relative to the nearness of a sensor target to the white target?

6. In general, do the criticisms of the statistics, or lack of statistics, used to analyze the data for the previous measures of performance also apply to the statistical analysis of this measure of performance?

7. After careful consideration of the preceding questions, are the conclusions reached in the report valid based on the analysis used?

III. CONCLUSION OF THE EXPERIMENT'S REPORT

The final conclusion of the report as to the overall fire effectiveness of mounted versus dismounted is as follows:

> Fire effectiveness of assaulting riflemen and machine-gunners mounted in armored personnel carriers is superior to that of dismounted riflemen and machine-gunners in the types of terrain used in this experiment...[17]

The use of statistics to reach an objective final conclusion is

---

*This is not an unreasonable assumption. Compare sensor target hits with total rounds expended in Table VI, p. 56, and consider that the 10 x 15 meter "window" is a good-sized target, even at the experiment's 200-meter maximum range.

[17] Ibid. IV, d., p. 15.

non-existent in the report. This exceedingly strong conclusion is somehow based on the findings discussed in section II of this chapter. The report simply states the above conclusion with no mention of the rationale employed to reach it.

According to the report, were not dismounted riflemen more accurate and did they not have a higher index of target area hits? Without so stating, apparently the report writers felt that dismounted inferiorities in the other two measures of performance overshadowed superiorities in the accuracy and index measures of performance. One can only conclude that subjectivity, and not statistical objectivity, led to the final conclusion of the report.

An objective conclusion that could have been inferred from the results of this experiment is that there is an indication that mounted firers actually can hit a target with some degree of accuracy.

Reconsider the measures of performance used in the experiment to describe fire effectiveness. Accuracy, or hit probabilities, can be granted as being a rather good indication of the "ability to defeat a point target". That measure of performance and its quantification is the only one in the experiment which is meaningful itself and meaningfully quantified by data. Time to obtain a first hit could have been meaningful had the experiment captured the proper data. Suppression is a concept which is extremely difficult to quantify and measure. So much depends on the psychological make-up of the enemy that it may not be possible to measure suppression at all.

Is an enemy better suppressed by weapons with a high rate of fire? Would the rate of advance of dismounted versus mounted troops influence the enemy's desire to expose himself and return defensive fire? The appropriateness of the measures of performance used to describe suppression is certainly questionable. Even if the measures were appropriate, was appropriate data used to quantify them? The subject troops were instructed to fire at a point target; they were not instructed to place suppressive fire, which is normally thought of as area fire, into the target areas. To say that the two types of fire are the same in this experiment is to assert that one white target in each target group properly constitutes a typical enemy defensive threat.

Time, apparently, was an overriding constraint on the conduct of this experiment. It appears that the experiment was fielded with insufficient planning as to how fire effectiveness should be quantified and measured in the eight or less days available to conduct the experiment in the field. It almost seems as though some of the measures of performance might have been defined, after the experiment was conducted, to fit the data captured in the field trials.*

Had additional troops been available, it should have been feasible to replicate six groups per mode in the same time it took to repeat trials three times on each of two groups per mode. This,

*The reader is invited to review the quote on page 26 of Chapter III.

however, would have required 72 riflemen if the 6-man group was retained and 36 machinegunners if the 3-man group was retained. Unless the data collection plans were changed, the data captured would still relate well only to accuracy, but at least 6 replications might have been sufficient to permit the use of a meaningful hypothesis test on the accuracy of mounted versus dismounted groups.

IV. SUMMARY OF COMMENTS ABOUT THE EXPERIMENT

Hypothesis testing should have been used to determine if differences existed in the quantified measures of performance. Operational differences should have been defined and sufficient replications conducted to permit an objective analysis using hypothesis testing with reasonably low levels of Type I and II error probabilities.

Blind comparisons of estimates can often lead to erroneous conclusions which are unsupportable when subjected to rigorous statistical examination. Remember that the omission of a confidence interval about a point estimate infers an exactness about the value of the estimate that does not exist. For example, if the experiment's results, say the hit probabilities, are to be used in a war game computer simulation, the extreme ends of the estimate's confidence interval could be tested in the simulation to determine how sensitive the simulation's output is to possible errors in the input values of hit probabilities. Use of only the point estimate with a belief in the

exactness of its value could invalidate the results of such a simulation.

The selection of good measures of performance is an extremely difficult task. But, once measures of performance are defined, the experimental design must include a data collection plan which will insure that data relevant to the measures are captured by the experiment.

On the surface, this report appears to have an aura of scientific credibility. After all, it contains "facts" and figures derived from a scientific investigation called field experimentation. It fallaciously assures the report's reader that the data are "statistically valid" since a "grand total of 324 repetitions" were performed. The findings are strong and assertive, mounted is better than dismounted under the conditions tested, with no mention of a possibility that the findings might be in error. In actuality, the report attempts to conceal a very subjective analysis of a poorly designed experiment in a cloak of statistical objectivity. Certainly subjectivity can, and in some instances should, play an important role in the analysis of an experiment, but it should be identified as such when it is employed and not hidden behind a facade of psuedo-scientific objectivity just to lend credibility to the analysis. If resources are limited and constrain the collection of sufficient data to perform a truly objective statistical analysis, yet experimentation must be performed to meet an important requirement for information, then

the results generated from a subjective analysis of limited data should be labeled as having been derived from a subjective analysis based on limited data and the judgment and experience of the analyst.

The critique of this experiment should evidence the need for the military officer, whose duties are the planning, conduct, analysis, and evaluation of field experimentation, to understand the basic concepts of experimental statistics.

BIBLIOGRAPHY

Ehrenfeld, Sylvain and Littauer, Sebastian B. Introduction to Statistical Method. New York: McGraw-Hill Book Company, 1964.

Ostle, Bernard. Statistics in Research. Ames: The Iowa State University Press, 1964.

Comparison of Fire Effectiveness - Mounted vs Dismounted. Experimental Report, DDC AD4080390, Fort Ord, California: U. S. Army Combat Developments Command Experimentation Command, 1964.

Experimentation Manual. Fort Ord, California: U. S. Army Combat Developments Command Experimentation Command, 1968.

Methodology Notebook for Action Officers. Fort Belvoir, Virginia: U. S. Army Combat Developments Command, 1967.

APPENDIX A

STATISTICS SURVEY QUESTIONNAIRE

This Appendix contains the format and wording of the survey questionnaire used to survey the 35 officers of the United States Army Combat Developments Command Experimentation Center. The introductory, instructional, and question portions of the survey questionnaire are exactly as were given to the surveyed subject with the following changes made for the convenience of the reader of this thesis: Blank spaces for the replies to essay or "list" type questions have been eliminated and such questions compacted with each other. Second, the key to the correct replies to the matching section of the questionnaire has been inserted next to the appropriate terms. Third, terms of the matching section whose definitions can be found in the USACDCEC Experimentation Manual are noted with an asterisk.

SURVEY OF USACDCEC OFFICERS' KNOWLEDGE OF STATISTICS

Background

Captain D. Mikkelson, USA, student in Operations Research at the Naval Postgraduate School, is directing his master's thesis efforts toward a paper on basic statistics as applied to Army field experimentation. The paper is to be written specifically for the Army officer who is connected with field experimentation so that he will have a better understanding of some of the important statistical concepts of experimentation.

The main purpose of this survey is to provide an insight to the statistical proficiency of the CDCEC officers directly connected with experimentation. The results of the survey should indicate the level of technical statistical language in which the paper should be written to be understandable and informative to you, the intended reader.

The questionnaire will not be simple to complete. To be at all thorough in its completion, each officer will probably expend at least one and one-half to two hours on the questionnaire. Do not become discouraged if difficulties are encountered in answering the questions; the questionnaire was intentionally designed to be difficult for officers with little or no knowledge of statistics.

Anonymityof officers completing the questionnaire will be preserved. No attempt will be made to isolate individual performance by name.

Your conscientious efforts in the completion of the questionnaire are appreciated.

Questionnaire Instructions

The questionnaire is composed of four sections: (1) general information, two parts, (2) essay questions and (3) matching terms with definitions. You may use pencil or pen of any color except red for noting your answers.

You are expected to work on the questionnaire individually, but you may use any written reference that you normally use in the performance of your duties. Please do not ask Scientific Support Laboratory personnel to assist you.

Administrative questions about the questionnaire should be directed to Lt. Col. Phillips, XT 4481.

QUESTIONNAIRE

General Information - Part I

1. Rank _____. 2. Branch _____. 3. Years on active duty _____.

4. Total time at CDCEC _____ months.

5. Current assignment _____. (e.g., G-3 FED, Team I)

6. Total time in current assignment _____ months.

7. Years of college attendance _____.

8. Degree(s) and field(s) _____.

9. Year that degree(s) was(were) granted _____.

10. Indicate below if you have had any formal instruction in statistics.

What type of instruction	Hours of instruction	When (yr)

11. Would having a better knowledge of experimental statistics assist you in performing your current duties more effectively?

Circle one - a. Yes, definitely; b. marginally valuable;

c. don't think so d. definitely not.

Essay Questions

If there is insufficient room for your reply, use the back of the page or add an extra page. Please attempt to make studied responses.

1. What is the use of STATISTICS in field experimentation?

2. Is there a difference between "field experimentation and "troop testing" of men and material? Explain your answer.

3. Suppose we want to collect experimental data by making observations on test subjects who are a sample from a large population. Is it better to record single observations on each of many different subjects or to record repeated observations on the same subject to produce the data? Explain your answer.

4. Assume we are comparing the performances of two configurationally different platoons in a certain measure of effectiveness. What is the meaning of the statement "there is a significant difference at the 5 per cent level between platoons" or "the significance level for a difference between platoons is 5 per cent"?

5. a. What, if any, "scientific" terms, phraseology, concepts, theories, etc. are used frequently in your dealings with the civilian scientist that you do not really understand?

b. Do you feel that there is a "communication gap" between the military and the civilian scientist? Explain your answer.

6. Would a reference on experimental statistics, written specifically for officers with little or no formal statistical training, be of some value to you? If yes, in what way?

Matching

This section requires you to select the numbered definition on pages Q-7 and Q-8 which is most closely related to the lettered term or phrase. Please do not guess. Some definitions may be used more than once for different terms or phrases. However, there is only one best definition for each term or phrase. Write the definition number to the left of the appropriate term. The first term has been completed as an example.

For your convenience, the pages containing the definitions may be detached from the questionnaire. When you have completed the matching, the pages may be discarded.

TERMS

(Key)	10	A. Random Sampling
11*	_____	B. Purpose of Replication in Experimentation
12	_____	C. Type I or Alpha Experimental Error
20	_____	D. Type II or Beta Experimental Error
12	_____	E. Producer's Risk
20	_____	F. Consumer's Risk

26	_____	G. Power of a Test
27	_____	H. Test Statistic
2*	_____	I. Variance or Standard Deviation of Observations
24	_____	J. Null Hypothesis
23	_____	K. Alternative Hypothesis
19	_____	L. Hypothesis Testing
22	_____	M. Operating Characteristic Curves
16*	_____	N. Sample Mean
17*	_____	O. Sample Median
18*	_____	P. Sample Mode
9*	_____	Q. Independent Experimental Variable
7*	_____	R. Dependent Experimental Variable
8*	_____	S. Uncontrolled Experimental Variable
4	_____	T. Statistical Significance
13*	_____	U. Confidence Interval

PLEASE GO BACK OVER THE PRECEDING THREE SECTIONS TO INSURE THAT YOU HAVE ANSWERED ALL QUESTIONS.

General Information - Part II

1. List any written references you used in completing the questionnaire.

2. Total time you devoted to the completion of the questionnaire.

_____ hrs _____ min.

*These terms are similarily defined in the glossary of the USACDCEC Experimental Manual.

DEFINITIONS

1. The probability of inferring erroneous conclusions from valid data.

2. A measure of the closeness to the average or grouping around the average that is exhibited by a series of similar data.

3. Provides more data so that the validity of the results of the experiment is increased.

4. The rejection of the null hypothesis at a particular level of Type I error.

5. Devices for determining true characteristics from experimental data.

6. Determining whether the null or the alternative hypothesis should be used in the design of an experiment.

7. A variable whose magnitude is expected to vary as a result of variation in the magnitude of another variable.

8. A variable whose fluctuation in magnitude will have little or no influence on the relationships between other variables.

9. A variable that is deliberately changed or allowed to change in magnitude in order to determine the effects of such change on other variables.

10. Permitting the "laws of Chance" to govern the selection of subjects on which observations are to be taken.

11. Serves to average out random sources of error and supplies an estimate of experimental error.

12. The probability of concluding that a significant difference exists when, in fact, there is no true significant difference.

13. An allowance for error around an estimate with a degree of belief in the size of the error; or the degree of belief that a certain range of values contains the true value being estimated.

14. The importance of taking accurate and precise measurements when performing experimentation.

15. A specific range of values around an experimental estimate in which the true value being estimated has a certain probability of

occurring; or an interval in which the true value being estimated has a certain probability of occurring, where the certain probability is greater than zero and less than one.

16. The sum of the values of data points divided by the number of data points summed.

17. The value about which 1/2 the data points have greater value and 1/2 the data points have less value.

18. The value occurring with the highest frequency.

19. Producing and analyzing a test statistic whereby the null hypothesis can be accepted or rejected.

20. The probability of concluding that no significant difference exists when, in fact, there is a true significant difference.

21. The probability of inferring valid conclusions from bad or erroneous data.

22. The relationship between Type I and Type II error and the number of replications in an experiment.

23. A statement of inequality about the true condition being tested.

24. A statement of equality about the true condition being tested.

25. The probability of concluding that no significant difference exists when, in fact, there is no true significant difference.

26. The probability of concluding that a significant difference exists when, in fact, there is a true significant difference.

27. An experimental estimate, computed from experimental data, of an actual condition that exists in nature which is used in hypothesis testing.

28. The result of reduction of raw data which is used to verify the reliability and validity of a test or experiment.

29. A measure of the degree of error that might occur in a series of similar observations.

30. I don't know the meaning and don't want to guess.

APPENDIX B

SUMMARY OF AND COMMENTS ON SURVEY RESULTS

The contents of this Appendix supplements the discussion of Chapter II with more detail. Familiarity with Appendix A is a prerequisite for reading this Appendix. Frequent reference will be made to the questions of the essay section of the survey questionnaire and to the terms and definitions of the matching section of the questionnaire. Hopefully, the use of numbers and letters in lieu of the complete wording of the essay questions, terms, and definitions will produce sufficient clarity and brevity in the discussion to compensate the reader for those instances where he may be inconvenienced by a need to refer to Appendix A.

General Information About the Survey Sample

The number of officers participating in the survey by rank were: 6 Lieutenant Colonels, 14 Majors, 1 Captain, 7 First Lieutenants, 6 Second Lieutenants, and 1 Warrant Officer. The number of surveyed officers by branch were: Armor-12, Artillery-13, Infantry-2, Corps of Engineers-2, Signal Corps-1, Ordnance Corps-3, Chemical Corps-1, Womens Army Corps-1, and Aviation-1.

One-half of the officers had completed 10 years or more of active duty. The average (mean) time with USACDCEC was 10.5 months. The average (mean) time of an officer in his present job assignment at USACDCEC was 6.8 months.

The thirty-five officers surveyed have completed a total of 142 years of college level education; an average of just over 4 years per officer. Twenty-seven officers have at least a bachelor's degree and 9 of the 27 have a master's degree. Twenty officers indicated that they have had no formal education or training in statistics.

To complete the questionnaire, fourteen officers indicated that they used at least one written reference; ten of those officers indicated they used the USACDCEC Experimentation Manual.

The average (mean) questionnaire completion time for thirty-one officers was 86.6 minutes; four did not indicate their completion time. The least and longest completion times were 30 and 220 minutes respectively. Fourteen took 75 minutes or less and seventeen took 80 minutes or more. Only three took less than 60 minutes.

Results of Essay Question Section

The answers to the first four essay questions* were evaluated in a subjective manner and given ratings of excellent, good, fair, or poor. (Poor includes instances where no answer was given). Table X presents the results of the ratings.

---

* Refer to page 82 of Appendix A for a list of essay questions.

TABLE X

DISTRIBUTION OF RATINGS RECEIVED BY QUESTION

Question Number	Excellent	Good	Fair	Poor
1.	7	13	10	5
2.	10	10	3	3
3.		16	13	6
4.	1	3	8	23

The following paragraphs are comments on the essay question replies. Recall that questions 4 and 6 were discussed in Chapter II.

Question One. Only a few officers seemed to be unable to convey fully that statistics lends scientific rigor and objectivity in the planning of experiments and in the analysis and interpretation of experimental data.

Question Two. The definitions of "field experimentation" versus "troop testing" found in the USACDCEC Experimentation Manual were the criteria for rating replies to this question.

Question Three. None of the subjects were rated excellent because of their failure to mention specifically the implications of estimating experimental error and variance as it applies to extrapolating results to the overall population. But, sixteen agreed it is generally best to replicate with new subjects.

Question Five. a. Sixteen officers felt that at least some of the terms and phrases used by the civilian scientist were not fully

understood. However, specific terms or phrases cited as examples by the officers did not recur with sufficient frequency to warrant mentioning. Eight officers indicated that they had no problems understanding the scientific language used. Eleven officers either gave no reply or were non-committal. b. Twenty-one officers felt that there was at least some degree of a "Communications Gap". But, as many pointed out, this is inherent in the soldier-scientist relationship. Some officers inferred that the scientist could do a better job of simplifying his explanations of questions from the military. Nine officers felt that there was no real "Communications Gap" and five did not reply or were non-committal.

Results of the Matching Section

Further insight to the nature of the responses to the matching section is gained by considering the replies to individual terms or phrases.* Only incorrect replies which occurred with high frequency for a certain term are discussed. The page references in parentheses following some of the terms are from the Glossary of the USACDCEC Experimentation Manual.

B. - Purpose of Replication in Experimentation (p. A-24). Twenty-seven subjects selected definition #3. Definition #3 was purposely inserted to draw incorrect responses. The key word in definition #3 is "validity". Validity of results is not directly

*Refer to pages 85 and 86 of Appendix A for the list of definitions.

dependent on replication. Results are valid dependent on what kind of a measure of performance is measured and how it is measured. Had "accuracy" or "confidence" been used in lieu of "validity", definition #3 would have been a good response.

U - Confidence Interval (p. A-6). Eighteen subjects selected definition #15. Definition #15 is a common misunderstanding about confidence intervals. Once a specific interval is derived from experimental data, either the true value being estimated is or is not contained in the interval with probability 0 or 1. A specific 95% confidence interval does not mean that the probability of the true value lying within the specific interval is 95%; it means that if the experiment were exactly repeated 100 times, 95 of the 100 intervals generated from the data would be expected to include the true value, thus the implication of "95% confidence" in any specific interval.

Six officers indicated that they did not know the definition.

Q - Independent Experimental Variable (p. A-28). Nine subjects selected definition #8, a definition of an uncontrolled experimental variable.

Six subjects indicated that they did not know the definition.

R - Dependent Experimental Variable (p. A-29). Four subjects selected definition #9, a definition of an independent experimental variable.

Six subjects indicated that they did not know the definition.

understood. However, specific terms or phrases cited as examples by the officers did not recur with sufficient frequency to warrant mentioning. Eight officers indicated that they had no problems understanding the scientific language used. Eleven officers either gave no reply or were non-committal. b. Twenty-one officers felt that there was at least some degree of a "Communications Gap". But, as many pointed out, this is inherent in the soldier-scientist relationship. Some officers inferred that the scientist could do a better job of simplifying his explanations of questions from the military. Nine officers felt that there was no real "Communications Gap" and five did not reply or were non-committal.

Results of the Matching Section

Further insight to the nature of the responses to the matching section is gained by considering the replies to individual terms or phrases.* Only incorrect replies which occurred with high frequency for a certain term are discussed. The page references in parentheses following some of the terms are from the Glossary of the USACDCEC Experimentation Manual.

B. - Purpose of Replication in Experimentation (p. A-24). Twenty-seven subjects selected definition #3. Definition #3 was purposely inserted to draw incorrect responses. The key word in definition #3 is "validity". Validity of results is not directly

*Refer to pages 85 and 86 of Appendix A for the list of definitions.

dependent on replication. Results are valid dependent on what kind of a measure of performance is measured and how it is measured. Had "accuracy" or "confidence" been used in lieu of "validity", definition #3 would have been a good response.

U - Confidence Interval (p. A-6). Eighteen subjects selected definition #15. Definition #15 is a common misunderstanding about confidence intervals. Once a specific interval is derived from experimental data, either the true value being estimated is or is not contained in the interval with probability 0 or 1. A specific 95% confidence interval does not mean that the probability of the true value lying within the specific interval is 95%; it means that if the experiment were exactly repeated 100 times, 95 of the 100 intervals generated from the data would be expected to include the true value, thus the implication of "95% confidence" in any specific interval.

Six officers indicated that they did not know the definition.

Q - Independent Experimental Variable (p. A-28). Nine subjects selected definition #8, a definition of an uncontrolled experimental variable.

Six subjects indicated that they did not know the definition.

R - Dependent Experimental Variable (p. A-29). Four subjects selected definition #9, a definition of an independent experimental variable.

Six subjects indicated that they did not know the definition.

S - Uncontrolled Experimental Variable (p. A-29). Five subjects selected definition #9. Seven subjects indicated that they did not know the definition.

N, O, P - Sample Mean, Median, Mode (p. A2, 17, 18). These terms posed no problem for most subjects; 38% of all correct answers were on these three terms. One officer properly used definition #27 for N, sample mean, but definition #16 was considered the "best" answer.

L - Hypothesis Testing. Eleven subjects chose definition #6. Definition #6 is certainly not correct for term L. It was purposely inserted in the definition list because it "sounded" correct.

M - Operating Characteristic Curves. Eight subjects chose definition #5, another definition like #6, inserted because it "sounded" correct.

T - Statistical Significance. There were only two correct replies. Ten subjects replied that they did not know the definition. The remaining twenty-three subjects selected 14 different definitions from the list of definitions for their replies.

G - Power of a Test. There was only one correct reply. Sixteen subjects replied that they did not know the definition. The remaining eighteen subjects selected 11 different definitions for their replies.

An interpretation of the information in Table XI and the nature of replies to terms G, L, M, and T could imply that guessing may

TABLE XI

TYPES OF REPLIES TO SELECTED TERMS

Term	1 Wrong Reply (guess?)	2 Didn't Know, #30 (no guess)	Ratio 1/2
T	23	10	2.30
L	17	8	2.13
I	12	6	2.00
M	21	12	1.75
H	14	12	1.17
G	18	16	1.12
E	13	19	0.68
K	12	19	0.63
F	10	20	0.50
J	10	20	0.50
C	7	24	0.28
D	4	26	0.15
Totals	163	192	0.85

have been prevalent, at least on certain terms. Intuitively, one would expect the "guess" to "no guess" ratio to be somewhat less than one since the questionnaire instructions for the matching section specifically asked subjects not to guess. The effect of guessing in the results would be the artificial inflation of the number of terms correctly identified by certain, though perhaps few, "lucky" or chance choices of correct definitions. The surveyed officers, therefore, might actually know less about the terms of the matching section than the results indicate. Notice that terms B and U are properly omitted from this analysis since definitions #3 and #15 were not included specifically to draw guesses.

However, the "guess" to "no guess" ratios for terms C, D, E, F, J, and K were less than one. The nature of these terms and apparent lack of understanding of them may have discouraged guessing. Terms N, O, P, Q, R, and S were not analyzed for guessing since each term was answered correctly in more than 50% of the replies and has been previously discussed.

The overall response to the 9 terms found in the USACDCEC Experimentation Manual produced 181 correct replies out of 315 opportunities (9 x 35), or 57.5% correct, with an average of 5.2 correct replies per officer.

Recalling that twenty-five officers indicated that they did not use the manual when completing the questionnaire, the response data was analyzed to attempt to determine if those who did use the manual

performed better than those who did not. Surprisingly, there was no evidence to clearly indicate that correct responsiveness was enhanced by the use of the manual. As examples to the contrary, the three officers who scored well, 13, 13, and 15 correct, indicated that they did not use the manual. One officer who did use the manual scored only 3 correct.

Finally, consider the response to the 11 terms not found in the manual. There were 50 correct replies out of 385 opportunities (11 x 35), or 13% correct, with an average of 1.40 correct replies per officer. It is most interesting to note that the five officers of Team V contributed 30 of the 50 correct replies giving them 55% correct, an average of 6.00 correct per officer, as compared with 6% correct, an average of 0.67 correct per officer for the survey sample minus Team V officers. Four of the five Team V officers surveyed had three or more college level hours of instruction in statistics.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	20
2.	Library Naval Postgraduate School Monterey, California 93940	2
3.	Operations Analysis Department Naval Postgraduate School Monterey, California 93940	1
4.	Commanding General U. S. Army Combat Developments Command Experimentation Center Fort Ord, California 93941	15
5.	Professor J. Bryce Tysver Department of Operations Analysis Naval Postgraduate School Monterey, California 93940	1
6.	Professor James G. Taylor Department of Operations Analysis Naval Postgraduate School Monterey, California 93940	1
7.	Captain David W. Mikkelson, USA % A. L. Mikkelson Wakonda, South Dakota 57073	1

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (*Corporate author*)

Naval Postgraduate School
Monterey, California 93940

2a. REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b. GROUP

3. REPORT TITLE

STATISTICS AND THE MILITARY OFFICER IN COMBAT DEVELOPMENTS EXPERIMENTATION

4. DESCRIPTIVE NOTES (*Type of report and, inclusive dates*)

Thesis - December 1968

5. AUTHOR(S) (*First name, middle initial, last name*)

David W. MIKKELSON

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
December 1968	97	5

8a. CONTRACT OR GRANT NO	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO	
c.	9b. OTHER REPORT NO(S) (*Any other numbers that may be assigned this report*)
d.	

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Naval Postgraduate School Monterey, California 93940

13. ABSTRACT

The duty performance of military officers whose duties are the planning, conduct, analysis, and evaluation of field experimentation can be improved through a better understanding of experimental statistics. The role of statistics in the field experimentation conducted by the U. S. Army Combat Developments Command Experimentation Center typifies the role of statistics in military field experimentation. Selected officers of USACDCEC were surveyed to determine their understanding of some of the more important concepts of experimental statistics. The survey results indicate that most of these officers lack a basic knowledge of experimental statistics. Based on insights gained from the survey, statistical training of certain USACDCEC officers is recommended. Statistical concepts not well understood by the surveyed officers are defined and discussed. A field experiment conducted by USACDCEC is used to exemplify the applications of statistical techniques and the use of measures of performance in field experimentation.

DD FORM 1 NOV 65 1473 (PAGE 1)
S/N 0101-807-6811

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Statistics						
Field experimentation						
Statistical concepts						
Combat developments						